DOCUMENT RESUME

ED 271 492 TM 860 400

AUTHOR Livingston, Samuel A.
TITLE Adjusting Scores on Examinations Offering a Choice of Questions.
PUB DATE Apr 86
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; *Error of Measurement; Essay Tests; Mathematical Models; Multiple Choice Tests; *Scoring Formulas; Statistical Analysis; Statistical Studies; *Testing Problems; *Test Items; Test Theory

ABSTRACT

This paper deals with test fairness regarding a test consisting of two parts: (1) a "common" section, taken by all students; and (2) a "variable" section, in which some students may answer a different set of questions from other students. For example, a test taken by several thousand students each year contains a common multiple-choice portion and a common essay portion but also a variable essay portion, in which the test-taker may choose to answer any one of five questions. On this test the questions that the test-taker may choose from are intended to be of equal difficulty. When the scoring has been completed and the results tabulated, the data occasionally suggest that two or more essay questions may not have been of equal difficulty. If there had been no reason to believe, a priori, that the questions on the variable portion were of equal difficulty, the scores would need to be adjusted in such a situation. Problems arise with the two adjustments: option A leads farthest away from the assumption of equal difficulty when the evidence against it is weakest; and option B is equivalent to assuming that the questions in the variable portion are, in fact, equally difficult. A compromise is proposed that is closer to option A when the common portion predicts the variable portion accurately and closer to B when it does not. (PN)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED271492

February 27, 1986

# Adjusting Scores on Examinations Offering a Choice of Questions

Samuel A. Livingston
Educational Testing Service

A paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April, 1986.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S A. Livingston

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 400

# Adjusting Scores on Examinations Offering a Choice of Questions

Samuel A. Livingston
Educational Testing Service

This paper is about fairness in testing. It deals with a particular type of test. This type of test consists of two or more parts. At least one part is a "common" section, taken by all the students. But the test also contains at least one "variable" part, in which some students may answer a different set of questions from other students. Often these tests allow the student a choice of questions on the variable portion of the test. For example, one test taken by several thousand students each year contains a common multiple-choice portion and a common essay portion but also a variable essay portion, in which the test-taker may choose to answer any one of five questions. On this test--and possibly on other tests that allow a choice of questions--the questions that the test-taker may choose from are intended to be of equal difficulty. In fact, the developers of the test work very hard to produce questions of comparable difficulty, and the scoring leaders work very hard to establish and maintain scoring standards that are comparable across questions.

Nevertheless, when the scoring has been completed and the results tabulated, the data occasionally suggest that two or more essay questions may not have been of equal difficulty. Consider the example in Table 1. A comparison of the groups answering questions 5 and 6 should cause us at least to question an assumption of equal difficulty. Group 5, on the basis of the common portions, appears to be as able as the other groups, but their scores on the variable portion average a third of a standard deviation lower. Group 6 appears, on the basis of the common portions, to be somewhat weaker than the other groups, but their scores on the variable portion average slightly higher.

If we had no reason to believe, a priori, that the questions on the variable portion were of equal difficulty, we would surely want to adjust the scores in such a situation. We would assume that groups of students whose performance on the common portion indicates they are of equal ability should also receive similar scores on the variable portion. Probably the simplest way to make an adjustment based on this assumption would be to estimate a "question effect" for each question and subtract this estimated "question effect" from the student's score on the variable portion. This kind of an adjustment would completely disregard all the attempts to make the questions on the variable portion equally difficult.

Of course, we could use a much more sophisticated type of adjustment. For example, instead of conditioning on the total score from the common portion, we could condition on some combination of subscores. Or we could condition on a weighted composite of the items in the common portion, choosing weights that maximize the difference between the groups of students choosing different questions on the variable portion of the test. Instead of estimating a constant question effect, we could adjust for differences in the conditional means and the conditional standard deviations, and maybe some higher moments of the conditional distributions. Stating this approach as generally as possible, we would condition on some function of the response pattern from the common portion, which would serve as a common measure of ability. Then we would assume that some characteristics of the distribution of scores on any given question in the variable portion would be the same, in some specified way, for all groups of students of equal ability, as indicated by the common portion. In particular, we would assume that the scores on any question in the variable portion would have been the

same for the students who did not answer the question as they were for the students who did answer it, when we condition on the common ability measure. But even with this very flexible approach, we would still be disregarding all the attempts to create questions of equal difficulty on the variable portion.

This approach is presented in Table 2 as "Option A". We cannot observe the responses of Group 2 to question 1. If Group 2 had answered question 1, how would they have performed, in comparison to Group 1? To answer this question, we condition on the common portion and then assume that no further ability difference exists between the two groups. What might make us uncomfortable about such an approach? Consider the case in which the responses to the common portion do not do very well at predicting scores on the questions in the variable portion. Figure 1 presents a very simple example, using just the total score on the common portion as the predictor. The solid ellipses represent the data we can observe; the dashed ellipses represent the distributions we impute under this assumption. You can see how the assumption implies that Group 2, with much lower scores on the common portion, would have done nearly as well as Group 1, if they had taken question 1.

There is another problem with the conditionally-equal-ability assumption of Option A. If the relationship between the common portion and the variable portion is weak, the reason may well be that the two portions measure somewhat different skills. Yet this is exactly the case in which the imputed score distribution for Group 2 on question 1 will be <u>farthest</u> from their actual score distribution on question 2. That is, a <u>weak</u> relationship with the covariate leads to a <u>large</u> adjustment. Remember, we

have some non-statistical information telling us that the questions are at least approximately equal in difficulty. The traditional approach of Option A leads us farthest away from the assumption of equal difficulty when the evidence against it is weakest.

If Option A is not fully satisfactory, what about Option B? Option B says to assume that if Group 2 had taken question 1, they would have done just as well on Question 1 as they actually did on Question 2. This assumption is equivalent to assuming that the questions on the variable portion are, in fact, equally difficult. Under Option B, we would never adjust the scores on the variable portion, no matter what the scores on the common portion looked like. This assumption might not make us too uncomfortable in a situation like that of Figure 1, but look at Figure 2. In this example, the scores on the common portion are strongly related to scores on the variable portion. Yet, Group 2, with much lower scores on the common portion, gets much higher scores on the variable portion.

What we need is some sort of compromise between the two approaches I have labeled Option A and Option B, preferably a compromise that depends on the data. We would like a solution that is closer to Option A when the common portion predicts the variable portion accurately and closer to Option B when it does not. The only solution I have been able to come up with is one that requires a subjective decision. This approach says: Look at the difference between conditional means, compare it with the size of the conditional standard deviation, and ask yourself, "How big a difference am I willing to believe is a genuine difference in ability between the groups?" Option A, the basic covariance adjustment, says "None--any difference in conditional means must be the result of differences in question difficulty

(or scoring standards, etc.)." Option B, which leads to no adjustment, says, "All of it--any difference I observe must be a genuine ability difference, even though we are comparing students who are equal on x." I say, why force yourself to choose one or the other of these extreme positions. Why not say, "I will believe that a difference up to one, or two, or three conditional standard deviations could be due to genuine ability differences. I will adjust so as to remove any difference beyond that."

This approach does have the property of producing an adjustment that is larger when x predicts y more accurately. The more accurate the prediction, the smaller the conditional standard deviation, and the smaller the allowable difference between conditional means.

There is one feature of an adjustment based on this approach that runs counter to most people's idea of fairness, but we can correct the problem with a small modification. The problem is this: Suppose we apply the principle strictly, adjusting away any differences beyond the amount we have specified in terms of the conditional standard deviation. Then we could have a situation, in one of the groups, where two students could have the same unadjusted y score, but the student with the higher x score could receive a lower adjusted y score. To prevent this kind of unfairness we can introduce an additional constraint: the size of the adjustment, that is, the number of points to be added to or subtracted from a student's y score, must be the same for all students answering the same question on the variable portion. The resulting adjustment would take the form of a constant for each group, to be added to (or subtracted from) the Y scores of all students in the group.

In practice, the adjustment might be based on a much simpler model, treating the regression of Y on x in each group as linear and homoscedastic. Table 3 shows the equations that describe this adjustment in terms of the observed x and y scores. We would regress Y on x in each group to get an equation for y-hat, the conditional mean, and an estimate of the residual standard deviation. We would then compute a pooled regression equation for y-hat, weighting each of the group expressions by the number of students in the group. Finally, we would compute, for each group, the difference between the group y-hat and the pooled y-hat, divided by the residual standard deviation for that group. If the absolute value of this number were smaller than the value we specified as the biggest difference we would believe, we would make no adjustment. If it were larger than the specified value, we would subtract off the specified value, and the remainder would be the size of the adjustment we would make to the score of each student in the group.

What makes the problem of adjusting for different questions unlike the problem of adjusting for different readers? Certainly there are similarities. In both cases there has been a lot of effort to make adjustments unnecessary, and yet the data may suggest that there is still room for improvement. Just as different questions may measure different knowledge and skills, readers may differ in the types of knowledge they consider most important. But there is one important difference between the two situations. Papers are assigned to readers by a process which can be assumed to be approximately random with respect to students' ability. Therefore, it is perfectly reasonable to assume that the conditional distributions of essay scores-- conditional on some other part of the

test--should not differ systematically from one reader to another, beyond what we might expect from sampling variation. But when students are allowed to choose their own questions to answer, there could very well be systematic differences in the ability measured by the variable portion, even when we condition on the common portion. The question is how large an ability difference we are willing to believe is genuine. Statistics cannot answer this question for us, but they can give us a way to express our answer and translate it into a score adjustment that is consistent with what we believe.

Table 1. Example of Data from Common and Variable Portions of an Examination: Deviation of Each Group Mean from Combined Mean, in Terms of Combined Standard Deviation.

| | Group Selecting Variable Question | | | | |
|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 |
| Common multiple-choice portion | -0.12 | +0.03 | +0.04 | +0.10 | -0.35 |
| Common essay portion | -.029 | +0.05 | +0.18 | 0.00 | -0.20 |
| Variable essay portion | -0.12 | +0.08 | -0.03 | -0.34 | +0.14 |
| Number of students | 3,411 | 38,445 | 1,390 | 10,382 | 5,180 |

Table 2. Two possible assumptions.

Let $Y_1$ = score on variable question 1

$Y_2$ = score on variable question 2

$\underline{x}$ = vector of responses on common portion

$F_1$ = distribution in group taking variable question 1

$F_2$ = distribution in group taking variable question 2

| | Question 1 | Ouestion 2 |
|---|---|---|
| Group 1 | observed $F_1(Y_1 \mid \underline{x})$ | unobserved $F_1(Y_2 \mid \underline{x})$ |
| Group 2 | unobserved $F_2(Y_1 \mid \underline{x})$ | observed $F_2(Y_2 \mid \underline{x})$ |

Option A: Assume $F_1(Y_2 \mid \underline{x})$ (unobserved) $= F_2(Y_2 \mid \underline{x})$ (observed)

$F_2(Y_1 \mid \underline{x})$ (unobserved) $= F_1(Y_1 \mid \underline{x})$ (observed)

Implies that, conditional on $\underline{x}$, groups 1 and 2 are equally able.

Option B: Assume $F_1(Y_2 \mid \underline{x})$ (unobserved) $= F_1(Y_1 \mid \underline{x})$ (observed)

$F_2(Y_1 \mid \underline{x})$ (unobserved) $= F_2(Y_2 \mid \underline{x})$ (observed)

Implies that questions 1 and 2 are equally difficult.

Table 3. Proposed solution:

1. In each group, regress $Y_i$ on x to get $\hat{y}_i \mid x = a_i + b_i x$ and an estimate of the residual standard deviation $s(y_i \cdot x)$.

2. Weighting each expression for $\hat{y}_i$ by the number of students in the group, compute a pooled regression equation

$$\hat{y}_{pooled} \mid x = \sum n_i(a_i + b_i x)]/[\sum n_i]$$
$$= a_{pooled} + b_{pooled} x$$

3. For each group, compute a standardized difference index at the group mean x score:

$$d_i = \frac{(\hat{y}_i \mid x_i) - \hat{y}_{pooled} \mid x_i}{s(y_i \cdot x)}$$

4. Let d* represent the maximum allowable standardized difference.

   If $|d_i| \leq d^*$, make no adjustment

   If $d_i > d^*$, let the adjusted $y_i$ be $y_i - (d_i - d^*)$.

   If $d_i < -d^*$, let the adjusted $y_i$ be $y_i + (|d_i| - d^*)$.

Figure 1. Option A: Covariance adjustment.
(Hypothetical example)

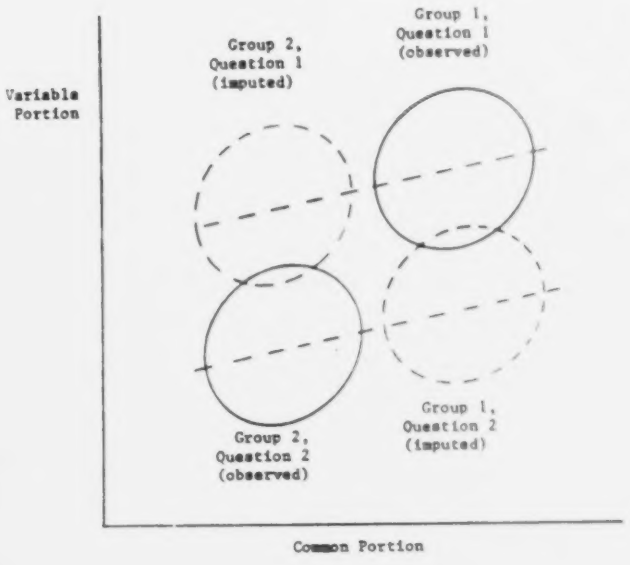
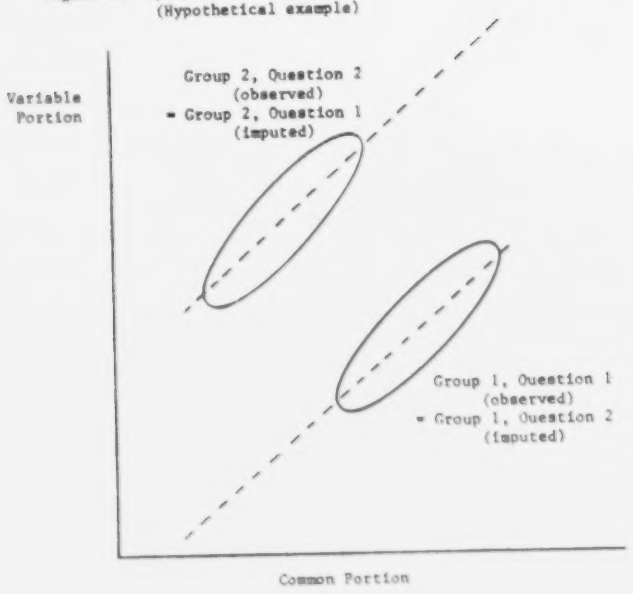

Variable Portion
Group 2, Question 1 (imputed)
Group 1, Question 1 (observed)
Group 2, Question 2 (observed)
Group 1, Question 2 (imputed)
Common Portion

Figure 2. Option B: No adjustment.
(Hypothetical example)


Variable Portion
Group 2, Question 2 (observed) = Group 2, Question 1 (imputed)
Group 1, Question 1 (observed) = Group 1, Question 2 (imputed)
Common Portion

BEST COPY AVAILABLE